

## AUTOMATED IMAGE INTERPRETATION AND RETRIEVAL SYSTEM

### Field of Invention

The present invention relates to the analysis of digital image and video signals for automated content interpretation in terms of high-level semantic labels called  
5 stereotypes. The labels can be subsequently used as a basis for tasks such as content-based retrieval and video abstract generation.

### Background

Digital image signal interpretation is the process of understanding the content of an image through the identification of significant objects or regions in the image and  
10 analysing their spatial arrangement. Traditionally the task of image interpretation required human analysis. This is expensive and time consuming. Consequently considerable research has been directed towards constructing automated image interpretation systems.

Most existing image interpretation systems involve low-level and high-level  
15 processing. Typically, low-level processing involves the transformation of an image from an array of pixel intensities into a set of spatially related image primitives, such as edges and regions. Various features can then be extracted from the primitives (e.g., average pixel intensities). In high-level processing, image domain knowledge and feature measurements are used to assign object or region labels, or interpretations, to the  
20 primitives and hence construct a description as to "what is present in the image".

Early attempts at image interpretation were based on classifying isolated primitives into a finite number of object classes according to their feature measurements. The success of this approach was limited by the erroneous or incomplete results that often result from low-level processing and feature measurement errors that result from the  
25 presence of noise in the image. Most recent techniques incorporate spatial constraints in the high-level processing. This means that ambiguous regions or objects can often be recognised as the result of successful recognition of neighbouring regions or objects.

More recently, the spatial dependence of region labels for an image has been modelled using statistical methods, such as Markov Random Fields (MRFs). The main  
30 advantage of the MRF model is that it provides a general and natural model for the interaction between spatially related random variables, and there are relatively flexible optimisation algorithms that can be used to find the (globally) optimal realisation of the

field. Typically the MRF is defined on a graph of segmented regions, commonly called a Region Adjacency Graph (RAG). The segmented regions can be generated by one of many available region-based image segmentation methods. The MRF model provides a powerful mechanism for incorporating knowledge about the spatial dependence of semantic labels with the dependence of the labels on measurements (low-level features) from the image.

Digital video is generally assumed to be a signal representing the time evolution of a visual scene. This signal is typically encoded along with associated audio information (e.g., in the MPEG-2 audiovisual coding format). In some cases information about the scene, or the capture of the scene, is also encoded with the video and audio signals. The digital video is typically represented by a sequence of still digital images, or frames, where each digital image usually consists of a set of pixel intensities for a multiplicity of colour channels (eg., R, G, B). This representation is due, in a large part, to the grid-based manner in which visual scenes are sensed.

The visual, and any associated audio signals, are often mutually correlated in the sense that information about the content of the visual signal can be found in the audio signal and vice-versa. This correlation is explicitly recognised in more recent digital audiovisual coding formats, such as MPEG-4, where the units of coding are audiovisual objects having spatial and temporal localisation in a scene. Although this representation of audiovisual information is more attuned to the usage of the digital material, the visual component of natural scenes is still typically captured using grid-based sensing techniques (ie., digital images are sensed at a frame rate defined by the capture device). Thus the process of digital video interpretation remains typically based on that of digital image interpretation and is usually considered in isolation from the associated audio information.

Digital audio signal interpretation is the process of understanding the content of an audio signal through the identification of words/phrases, or key sounds, and analysing their temporal arrangement. In general, investigations into digital audio analysis have concentrated on speech recognition because of the large number of potential applications for resultant technology. eg., natural language interfaces for computers and other electronic devices.

Hidden Markov Models are widely used for continuous speech recognition because of their inherent ability to incorporate the sequential and statistical character of a

digital speech signal. They provide a probabilistic framework for the modelling of a time-varying process in which units of speech (phonemes, or in some cases words) are represented as a time sequence through a set of states. Estimation of the transition probabilities between the states requires the analysis of a set of example audio signals for the unit of speech (ie., a training set). If the recognition process is required to be speaker independent then the training set must contain example audio signals from a range of speakers.

### Summary of the Invention

In accordance with one aspect of the present invention there is provided a method of classifying a digital image, said method comprising the steps of:

providing a region adjacency graph representing the digital image; and  
analysing said region adjacency graph for predetermined patterns and for each identified pattern selecting a classification of said digital image.

In accordance with another aspect of the present invention there is provided a method of classifying a digital image signal, said method comprising the steps of:

providing a labelled region adjacency graph representing at least part of the digital image signal;

providing plurality of classifications, for each of a plurality of patterns, wherein each said pattern comprises:

- i) a set of labelled regions; or
- ii) a set of labelled regions and corresponding adjacency information;

analysing said labelled region adjacency graph for the presence of predetermined patterns; and

for each pattern identified, selecting from said plurality of classifications a classification for the digital image.

In accordance with still another aspect of the present invention there is provided an apparatus for classifying a digital image, said apparatus comprising:

providing means for providing a region adjacency graph representing the digital image; and

analysing means for analysing said region adjacency graph for predetermined patterns and for each identified pattern selecting a classification of said digital image.

In accordance with still another aspect of the present invention there is provided a computer program product comprising a computer readable medium having a computer program recorded for classifying a digital image, said computer program product comprising:

5           providing module for providing a region adjacency graph representing the digital image; and

          analysing module for analysing said region adjacency graph for predetermined patterns and for each identified pattern selecting a classification of said digital image.

In accordance with still another aspect of the present invention there is provided  
10   an apparatus for classifying a digital image signal, said apparatus comprising:

          region adjacency graph providing means for providing a labelled region adjacency graph representing at least part of the digital image signal;

          classification providing means for providing a plurality of classifications, for each of a plurality of patterns, wherein each said pattern comprises:

- 15                   (i) a set of labelled regions; or  
                  (ii) a set of labelled regions and corresponding adjacency information; and

          analysing means for analysing said labelled region adjacency graph for the presence of predetermined patterns, wherein for each pattern identified, said classification  
20   providing means provides a classification for the digital image selecting from said plurality of classifications.

In accordance with still another aspect of the present invention there is provided a computer program product comprising a computer readable medium having a computer program recorded for classifying a digital image signal, said computer program product  
25   comprising:

          region adjacency graph providing module for providing a labelled region adjacency graph representing at least part of the digital image signal;

          classification providing module for providing a plurality of classifications, for each of a plurality of patterns, wherein each said pattern comprises:

- 30                   (i) a set of labelled regions; or  
                  (ii) a set of labelled regions and corresponding adjacency information; and

analysing module for analysing said labelled region adjacency graph for the presence of predetermined patterns, wherein for each pattern identified, said classification providing module provides a classification for the digital image selecting from said plurality of classifications.

5

### **Brief Description of the Drawings**

The embodiments of the invention are described hereinafter with reference to the drawings, in which:

Fig. 1. is a block diagram of a digital video interpretation system according to the preferred embodiment;

Fig. 2 illustrates the video segment analyser of Fig. 1 according to the preferred embodiment;

Fig. 3A and 3B illustrate a representative segmented image and a corresponding region adjacency graph (RAG), respectively, in accordance with the embodiments of the invention;

Fig. 4 illustrates a frame event analyser of Fig. 2 and having a single application domain;

Fig. 5 illustrates a hierarchical arrangement of stereotypes;

Fig. 6 illustrates an alternative frame event analyser to that of Fig. 2 and having multiple application domains;

Fig. 7 illustrates the selection of a temporal region of interest (ROI) for a particular analysis event;

Fig. 8 illustrates a preferred contextual analyser for use in the frame event analyser of Fig. 4 or Fig. 6;

Fig. 9 illustrates cliques associated with the RAG of Fig. 3B;

Fig. 10 is a block diagram of a representative computer for use with a digital video source, with which the embodiments of the invention may be practiced;

Fig. 11 is a block diagram of a representative digital video source, with which the embodiments of the invention may be practiced; and.

Fig. 12 illustrates the video segment analyser of Fig. 1 according to an alternate embodiment, which is optionally integrated in a digital video coding system;

### **Detailed Description**

1. Overview

The present invention relates to a method, apparatus and system for automatically generating a high-level description of a digital image or video signal that captures (important) semantic content of the digital signal in spatial and time domains. Such descriptions can subsequently be used for numerous purposes including content-based retrieval, browsing of video sequences or digital video abstraction. An embodiment of the invention is described in the following section using the example of a digital video signal. Similar embodiments exist for digital image signals.

A digital video signal is taken to be the visual signal recorded in a video capture device. The signal is generally, but not necessarily, generated from a two-dimensional array of sensors (e.g., CCD array) at a specified sampling rate with each sample being represented by a (video) frame. The analysis of the spatial and temporal content of this signal can benefit from a range of contextual information. In some cases this contextual information is implicit within the digital video signal (eg., motion), and in other cases the information can be available from other associated sources (eg., the associated audio signal, recorded camera parameters, other sensors apart from the generally used visual spectrum sensors). The extent of the available contextual information is much greater than that available to a still image analysis process and has the additional property of time evolution.

The time evolution of a digital image signal in a video signal can be used to improve the success of the image interpretation process of a digital video signal. For example, motion information can be used to assist in the detection of moving objects, such as people, in the digitally recorded scene. Motion can also be used to selectively group regions in an image frame as being part of the background for a scene. The process of interpreting, or understanding, digital video signals also benefits from the identification of audio elements (speech and non-speech) in the associated audio signal. For example, words identified in an audio signal are able to assist in the interpretation process. Also the audio signal associated with a wildlife documentary can contain the sounds made by various animals that can help identify the content of the video signal.

In the embodiments of the invention, a high-level description is generated at selected analysis events throughout the digital video signal. The high-level description is based on an assignment of various semantic labels to various regions that

are apparent at the analysis event. At each analysis event, the video frame centred on the analysis event is automatically spatially segmented into homogeneous regions. These regions and their spatial adjacency properties are represented by a Region Adjacency Graph (RAG). The probabilistic model is then applied to the RAG. The model  
5 incorporates feature measurements from the regions of the frame, contextual information from a Region of Interest (ROI) around the frame, and prior knowledge about the various semantic labels that can be associated with the regions of the RAG. These semantic labels (eg., "person", "sky", "water", "foliage", etc.) are taken from a list which has been typically constructed for an appropriate application domain (e.g., outdoor scenes,  
10 weddings, urban scenes, etc). Users of an image or video database can utilise these semantic labels for the purposes of retrieval. However, users will often want to search by even higher-level expressions (e.g., beach scenes, crowd scenes, etc.). These higher-level expressions can be termed stereotypes and represent classifications of the video frame or image. Stereotypes can be assigned to video frames, or images, by detecting patterns of  
15 region labels and (their) corresponding adjacency in the RAG. This assignment requires a predetermined set of patterns of region labels to be provided, wherein each pattern is associated with a stereotype.

At each analysis event, the contextual information is used to bias the prior probabilities of the semantic labels (hereinafter labels) in a selected appropriate  
20 application domain. The analysis performed at a given analysis event also depends on the previous analysis events. This dependence is typically greater if two analysis events are close together in the time domain. For example, within a video segment, it is likely, but not exclusively the case, that labels selected for regions at previous recent analysis events are more probable than labels that have not been selected in the description of the current  
25 section of digital video.

The digital video interpretation system can operate with a single application domain or multiple application domains. If multiple application domains are being used then contextual information can be used to determine the most probable application domain. The application domains can be narrow (ie., few labels) or broad (i.e., many  
30 possible labels). Narrow application domains are typically used if very specific and highly reliable region labelling is required. For example, in a security application, it can be

desirable to be able to identify regions that are associated with people and cars, but the identification of these objects might be required with high reliability.

In the following detailed description of the embodiments, numerous specific details such as video encoding techniques, sensor types, etc., are set forth to provide a more thorough description. It will be apparent, however, to one skilled in the art that the invention can be practiced without these specific details. In other instances, well-known features, such as video formats, audio formats, etc., have not been described in detail so as not to obscure the invention.

## 2. Digital Video Interpretation System of the Preferred Embodiment

Figure 1 illustrates a probabilistic digital video interpretation system 160 according to the preferred embodiment of the invention. The digital video interpretation system 160 comprises a video segmenter 120 and a video segment analyser 140. The system 160 processes digital video source output which has been generated from a digital video source 100. Preferably the digital video source 100 is a digital video camera. The video segmenter 120 is coupled between the digital video source 100 and the video segment analyser 140.

The digital video interpretation system 160 can optionally be internally or externally implemented in relation to the digital video source 100. When the digital video interpretation system 160 is located inside the digital video source 100 (eg., within a digital camera), the interpretation system 160 can readily make use of additional camera information without having to explicitly store this additional information with the video and audio signals that typically constitute the audiovisual signal. For example, camera motion information in a digital video camera can be used to assist motion analysis of the digital video signal 110A. Further, operator-gaze location can provide information about where key objects in a scene are located and focal information (or other range information) can be used to generate depth of field information which can be used to generate a depth axis for the RAG as shown in Fig. 3B.

The input to the digital video interpretation system 160 is the digital video source output 110 that is captured using a device such as a digital video camera. The digital video source output 110 is usually composed of a digital video signal 110A and a digital audio signal 110B. Additional information 110C about the recorded scene can also be



available depending on the capture device. Additional information 110C can include camera parameters (such as focus information, exposure details, operator-gaze location, etc.) and other sensor information (e.g., infrared sensing).

In the digital video interpretation system 160, the video segmenter 120 segments  
5 the digital video signal into temporal video segments or shots 130 provided at its output. The resulting video segments 130 produced by the video segmenter 120 are provided as input to the video segment analyser 140. The video segment analyser 140 produces a sequence of labelled RAGs for each video segment 130.

The video segment analyser 140 of Fig. 1 generates and then attempts to  
10 optimally label regions in a sequence of RAGs using one or more appropriate application domains. The resulting sequence of labelled RAGs 150 represents a description of the content of the digital video signal 110A and is referred to hereinafter as metadata.

Embodiments of the invention can also be implemented externally in relation to a digital video source as indicated by the general purpose computer 900 depicted in Fig. 10.  
15 Alternatively, the digital video interpretation system can be implemented internally within a digital video source 1000 as depicted in Fig. 11.

With reference to Fig. 10, the general purpose computer 900 is coupled to a remote digital video source 1000. The video interpretation system is implemented as software recorded on a computer readable medium that can be loaded into and carried out  
20 by the computer. The computer 900 comprises a computer module 902, video display monitor 904, and input devices 920, 922. The computer module 902 itself comprises at least one central processing unit 912, a memory unit 916 which typically includes random access memory (RAM) and read only memory (ROM), and input/output (I/O) interfaces 906, 908, 914 including a video interface 906. The I/O interface 908 enables the digital  
25 video source 1000 to be coupled with the computer module 902 and a pointing device such as mouse 922. The storage device 910 can include one or more of the following devices: a floppy disc, a hard disc drive, a CD-ROM drive, a magnetic tape drive, or similar non-volatile storage devices known to those skilled in the art. The components 906 to 916 of the computer 902 typically communicate via an interconnected bus 918 and  
30 in a manner which results in a usual mode of operation of a computer system 900 known to those in the relevant art. Examples of computer systems on which the embodiments of the invention can be practised include IBM PC/Ats and compatibles, Macintosh

Computers, SunSparcstations, or any of a number of computer systems well known to those in the art. The digital video source 1000 is preferably a digital camera that is capable of recording a video signal into storage (e.g.: memory, magnetic recording media, etc.) and additional information, for instance, infrared data that is associated with the video signal. The digital video signal and any associated additional (contextual) information can be downloaded to the computer 900 where the interpretation and labelling processes are performed in accordance with the embodiments of the invention.

Alternatively, the embodiment of the invention can be practiced internally within a digital video source 1000, which is preferably a digital video camera. The digital video source 1000 comprises video capture unit 1002 (e.g., including a charge couple device) for capturing images and having sensors and/or mechanisms for providing focal data and other settings of the video capture unit 1002. The digital video source 1000 can also include sensors 1004 for capturing audio information, ambient and/or environmental data, positioning data (e.g., GPS information) etc. The sensors 1004 and the video capture unit 1002 are connected to a central processing unit of the digital video source 1000, with which the embodiments of the invention can be practiced. The processing unit 1006 is coupled to memory 1008, communications port 1010, and a user interface unit 1012. The user interface unit 1012 allows the operator of the video source 1000 to specify numerous settings in operational modes of the digital video source 1000. For example, the digital video source operator can select different application domains (eg., Outdoor Scenes, Urban Scenes, Wedding Scenes, etc.) to use with the interpretation system. Application domains can be downloaded to the capture device, electronically or via a possible wireless link. The memory 1008 can comprise random access memory, read only memory, and/or non-volatile memory storage devices. Both data and processing instructions for operating the processing unit can be stored in the memory 1008. The communications port 1010 permits communications between the digital video source 1000 and external devices such as a computer 900 of Fig. 10. The communications port 1010 is capable of transmitting and receiving data and instructions to both the memory 1008 and the processing unit 1006.

### 3. Video Segmenter of the Preferred Embodiment

The video segmenter 120 segments the digital video signal into temporal video segments or shots 130. Information regarding the motion of pixels in a frame (implicit within the digital video signal 110A) and/or any other supporting information that is available in the digital audio signal 110B or other information 110C can be used to assist the segmentation by the video segmenter 120. Known video segmentation techniques can be implemented in the video segmenter without departing from the scope and spirit of the invention.

#### 4. Video Segment Analyser of the Preferred Embodiment

The video segment analyser 140 generates a sequence of labelled RAGs 150 for each video segment 130. The RAGs are preferably three-dimensional with range information being obtained from the digital video source 100 shown in Fig. 1. Each RAG consists of a set of disjoint regions and a set of edges connecting the regions. Regions that are located in the same X-Y plane are assumed to be coplanar. In contrast regions located in different Z planes are assumed to correspond to regions at different depths in the depicted scene. In general, the use of the depth axis (eg., Z-axis) in the RAG depends on the availability of information to indicate that a particular region is located at a different depth than one or more other regions. For example, the depth axis can be utilised in a digital video interpretation system 160 in which focal information, or depth information, is available to determine the depth of particular regions. However, the video segment analyser 140 can generate a sequence of labelled RAGs 150, without the aid of depth information by treating substantially all disjointed regions as being coplanar.

Fig. 2 shows the operation of the video segment analyser 140 according to the preferred embodiment. In block 200 an initial frame in a video segment 130 from the video segmenter 120 of Fig. 1 is selected for analysis. A frame event analyser 202 receives the selected frame and temporal region of interests (ROIs) for that frame, as described hereinafter with reference to Fig. 7, and generates a labelled RAG. Next, in block 204, the generated RAG is stored and a decision block 206 is used to determine if the end of the video segment 130 has been reached. If the end of the video is reached, that is the decision block 206 returns true (yes), video segment processing terminates in block 208. Otherwise the decision block 206 returns false (no), so a next frame to be analysed in the video segment 130 is retrieved and processing returns back to the frame

event analyser 202. Preferably, each frame of a video segment 130 is selected and analysed in real-time. However, in practice the selection of frames to be analysed depends upon the application in which the digital video interpretation system is used. For example, real-time performance is not always possible when analysing each frame in some devices, comprising the digital video interpretation system, in which case only predetermined frames of a video segment are selected for analysis.

Fig. 3B is an example of a three-dimensional RAG 310 for the spatially segmented frame 300 shown in Fig. 3A. The spatially segmented frame 300 contains nine regions named R1 to R9. The region R1 contains sky. The regions R2, R3, and R9 contain land and the region R8 is a road. The region R4 is a house-like structure, and the regions R5 and R6 are projecting structures in the house. To indicate depth in Fig. 3A, border of regions are indicated with several thicknesses. In particular, the thickness of the respective borders indicate the frontedness of depth in the Z axis. The RAG 310 indicates connected edges of regions R1 to R9 in the segmented frame 300. The regions R1, R2, R3, R7, R8 and R9 are all located at the same approximate depth (as indicated by solid edge lines) in the RAG 310 but at different X-Y positions. The region R1 is sequentially connected to regions R2, R8, R9 on the one hand and to regions R3 and R7 on the other hand. In turn, the region R4 has an edge with region R2, R3, R7 and R8, but has a different depth as indicated by dashed or broken edge lines. Finally, the regions R5 and R6 share an edge with the region R4 but at a different, parallel depth indicated by dotted edge lines. Thus, the dashed and dotted lines cross different Z-planes.

## 5. Frame Event Analyser

The functionality of the frame event analyser 202 of Fig. 2 is described in greater detail with reference to Fig. 4 and Fig. 6. The steps of the frame event analyser shown in Fig. 4 use a single application domain (eg., outdoor scenes). Such an application domain can contain the knowledge and functionality to label frame regions, such as sky, water, foliage, grass, road, people, etc.

In Fig. 4, the current frame and ROIs 400 for each of the information sources (e.g., the digital video signal, digital audio signal, etc.) are provided to a contextual analyser 410 that uses the ROIs 400. In addition to the contextual analyser 410, the frame

event analyser 202 of Fig. 2 comprises a frame segmenter 450, an adjustment unit 430 for adjusting prior probabilities of labels in an application, and a region analyser 470.

The contextual information 400 available in the ROIs is analysed by a contextual analyser 410. Since there are multiple sources of contextual information, the contextual  
5 analyser 410 typically includes more than one contextual analysing unit.

A more detailed illustration of the contextual analyser 410, is provided in Fig. 8 in relation to the adjusting unit 430 and the application domain 440. The contextual analyser 410 shown in Fig. 8 receives contextual information 400 for the frame event, which preferably comprises an audio ROI, a motion analysis ROI and/or an infrared  
10 spectral ROI. The contextual analyser 410 itself can include an audio analysing unit 710, a motion analysing unit 720, and an infrared analysing unit 730. The outputs produced by the contextual analyser 410 are used by the adjustment unit 430 to alter the prior probabilities of the labels in the application domain 440 used by the region analyser 470 of Fig. 4. The audio analysing unit 710 can achieve this result by recognising key words  
15 or phrases in the digital audio signal located in the audio signal ROI and then checking to see whether these key words/phrases suggest that any particular label(s) are more likely to occur in the frame than other labels. Other contextual analyser units (eg., 720, 730) can directly alter the prior probabilities of the labels.

In a further preferred embodiment having a frame event analyser 210 with a  
20 single application domain 440, the prior probabilities for the labels can be adjusted by the adjusting unit 430 on the basis of a list of key words/phrases 420 being stored for each label in the application domain 440 with a prior probability-weighting factor for each key word/phrase. The higher the probability-weighting factor is, the more likely that a region described by that label exists in the frame. Other contextual analysis results can also be  
25 provided to the adjusting unit 430 in addition to, or in place of, key words 420.

The frame segmenter 450 segments a frame into homogeneous regions using a region-based segmentation method. Typically, the segmentation method uses contextual information extracted from the ROIs of the different information sources (eg., video 110A and audio 110B) to assist the segmentation process. For example, motion vectors can  
30 assist in the differentiation of moving objects from a background. If focal information is available then this information can be used to estimate distances and therefore differentiate between different object or region planes in the frame. The result of the

segmentation process carried out by the frame segmenter 450 is a RAG 460, such as that shown in Fig. 3, which is provided as input to the region analyser 470. Preferably this RAG is three-dimensional. The other input to the region analyser 470 is the application domain 440 in which the prior probabilities of the labels may have been adjusted depending on the contextual information.

The probabilistic-model-based region analyser, 470, labels optimally the regions in the RAG using the appropriate application domain, 440. The resulting labelled RAG represents a description of the content of the frame, or metadata, that can be used for higher-level processes, such as content-based retrieval. Preferably the region analyser uses an MRF (probabilistic) model to produce the labelled RAG 480. The MRF model is described in detail hereafter.

The pattern of labels provided by the RAG can be utilised to generate stereotypes. A stereotype is a higher level expression which can be used for the purposes of retrieval on a more generic basis than is provided by the semantic labels of the RAG. The labelled RAG that results from MRF analysis provides a list of all the labels identified in the image/frame. Further, adjacency information is provided (eg., a sky region is next to/above a water region). Many stereotypes can be generated from this lower-level analysis. In one example, a beach scene can be classified as a scene which includes a sky region, a water region and a sand region. The stereotype specification can include a requirement such as a minimum size of these regions (e.g., the water must be larger than some minimum size), adjacency requirements (eg., the sky must be immediately above the water region) and/or just the existence of a set of labels in the RAG without any further limitations. Some simple examples of stereotypes are shown in Table 1.

**Table 1. Examples of stereotypes being derived from a RAG.**

Stereotype	RAG Requirements
Beach Scene	Preferably contains the labels sky, water and sand. The sky region must be immediately above the water region. The water region must be larger than some minimum size
Crowd Scene	Preferably contain more than, say, 10 people
Portrait	Preferably contain a single human face which has a minimum size with respect to the image/frame size
Rural Road Scene	Preferably contain a road and a green vegetation label (such as grass or foliage)

5 In a preferred embodiment, the stereotypes are represented by a simple association table and are assigned on the basis of the table in the final phase of the analysis. The classification of the stereotypes is represented in a hierarchical system so that the search/retrieval at any level of the hierarchy is possible. An example of such a hierarchy for outdoor scenes is shown in Fig. 5. Similar hierarchies are used for people  
10 containing scenes. The hierarchical path can be stored with the stereotype in a metadata object which is associated with the image/video object or accessed by reference.

The generated stereotypes can form the basis of an image retrieval system. Images can be retrieved on the basis of keyword (stereotype name) matching. Alternatively, the available stereotypes can be presented to the user in the form of icons.  
15 For example, a "portrait" icon shows just a single face taking up the area of the icon, whereas a "crowd" icon represents the possibility of many people/faces. The icons can be presented in the form of icon trees through which the user can navigate, where the icon trees represent the hierarchical arrangement of stereotypes. The user can select the icon representing the stereotype of the desired image(s), and a query can be generated for the

request. The query is processed and all images matching the query are retrieved and presented to the user. It should be clear to someone skilled in the art that the stereotype icons provide a convenient means for formulating a query to an image (or video) database. Furthermore, the image retrieval system can be implemented without the costly requirement of manual annotation, since the stereotypes are automatically generated by a digital image or digital video interpretation system.

The principles of the preferred method in generating stereotypes have been described with reference to semantic labels contained in the region adjacency graph. It is not intended that the present invention be limited to the use of semantic labels. For example, the RAG can contain label texture e.g. variance information based on simply the mean colour of the region or the colour variance (ie. texture) of the region. Many stereotypes could be generated by analysis of labelled RAGs where the labels do not necessarily reflect semantic meaning.

Turning to Fig. 6 there is shown a frame event analyser which can be described in a substantially similar manner to that of Fig. 4, excepting that the frame event analyser now has multiple application domains. In a frame event analyser having multiple application domains (ie., as depicted in Fig. 6), each application domain can contain a list of key words/phrases and the role of a selecting unit 530 can include the selection of an application domain to be used in the analysis. Thus, the selecting unit 530 selects a most probable application domain and preferably adjusts prior probabilities of labels in the select domain.

Referring to Fig. 7, there is shown a time-line 600 for a video segment. A current frame 601 of the video segment is analysed with reference to one or more regions of interest (ROIs) 602 extracted from available contextual information 603, audio information 604 (signal) and video information 605 (signal).

Temporal boundaries of the ROIs can vary with the type of contextual information (see Fig. 7). For example, contextual information such as camera parameters, can extend over a large temporal period, and possibly the entire video segment. In contrast, the ROI for the video signal can be much shorter, possibly several frames before and after the frame being currently analysed. ROIs are not necessarily centred over the current frame, as shown in Fig. 7. They can, for example, just include previous frames.



Mathematically a RAG is defined to be a graph  $\mathbf{G}$  which contains a set  $\mathbf{R}$  of disjoint regions and a set  $\mathbf{E}$  of edges connecting the regions;  $\mathbf{G} = \{\mathbf{R}, \mathbf{E}\}$ . Video frame interpretation seeks to optimally label the regions in  $\mathbf{G}$ . If an application domain consists of a set of  $p$  labels,  $\mathbf{L} = \{L_1, L_2, L_3, \dots, L_p\}$  with prior probabilities,  $\mathbf{Pr}_{\mathbf{L}} = \{Pr_{L1}, Pr_{L2}, Pr_{L3}, \dots, Pr_{Lp}\}$ , which have been biased by an analysis of the contextual information, then the interpretation process can be viewed as one of estimating the most probable set of labels on the graph  $\mathbf{G}$ .

If the graph  $\mathbf{G}$  consists of  $N$  disjoint regions, then let  $\mathbf{X} = \{X_1, X_2, X_3, \dots, X_N\}$  be a family of random variables on the RAG. That is,  $\mathbf{X}$  is a random field, where  $X_i$  is the random variable associated with  $R_i$ . The realisation  $x_i$  of  $X_i$  is a member of the set of labels,  $\mathbf{L}$ . A neighbourhood system  $\Gamma$  on  $\mathbf{G}$  is denoted by:

$$\Gamma = \{n(R_i); 1 \leq i \leq N\} \quad (1)$$

where  $n(R_i)$  is a subset of  $\mathbf{R}$  that contains neighbours of  $R_i$ . Preferably, a neighbourhood system for a region  $R_i$  is that region and all other regions which have some common boundary with  $R_i$ .

Further,  $\Omega$  is the set of all possible labelling configurations,  $\omega$  denotes a configuration in  $\Omega$ :

$$\Omega = \{\omega = \{x_1, x_2, x_3, \dots, x_N\} : x_i \in \mathbf{L}, 1 \leq i \leq N\} \quad (2)$$

Then  $\mathbf{X}$  is a MRF with respect to the neighbourhood system  $\Gamma$  if:

$$P(\mathbf{X} = \omega) > 0, \text{ for all realisations of } \mathbf{X};$$

$$P(X_i = x_i | X_j = x_j, R_j \neq R_i) = P(X_i = x_i | X_j = x_j, R_j \in n(R_i)). \quad (3)$$

An important feature of the MRF is that its joint probability density function,  $P(\mathbf{X} = \omega)$ , has a Gibbs distribution.:

$$P(\mathbf{X} = \omega) = Z^{-1} \exp[-U(\omega) / T], \quad (4)$$

where  $T$  is the temperature, and  $U(\omega)$  is the Gibbs energy function. The partition function  $Z$  is given as follows:

$$Z = \sum_{\omega} \exp[-U(\omega) / T]. \quad (5)$$

The energy function can be expressed using the concept of "cliques". A clique  $c$ , associated with the graph  $\mathbf{G}$ , is a subset of  $\mathbf{R}$  such that it contains either a single region or

several regions that are all neighbours of each other. The cliques for each region in the RAG depicted in Fig. 3B are listed in Fig. 9. The region R1 has associated cliques {R1}, {R1, R2}, and {R1, R3}, for example.

The set of cliques for the graph  $\mathbf{G}$  is denoted  $\mathbf{C}$ . A clique function  $V_c$  is a function with the property that  $V_c(\omega)$  depends on the  $x_i$  values (ie., labels) for which  $(i \in c)$ . Since a family of clique functions is called a potential,  $U(\omega)$  can be obtained by summing the clique functions for  $\mathbf{G}$ :

$$U(\omega) = \sum_{c \in \mathbf{C}} V_c(\omega) \quad (6)$$

Region-based feature measurements obtained from the frame and prior knowledge are incorporated into the clique functions  $V_c$ . The likelihood of a particular region label  $L_i$  given a set of region feature measurements can be estimated using various methods which could involve the use of a training set (e.g., neural networks) or can be based on empirical knowledge. Similarly, prior knowledge can be incorporated into the clique functions  $V_c$  in the form of constraints that can, but need not be, measurement-based. For example, the constraints can be of the form that label  $L_i$  and  $L_j$  cannot be adjacent (ie., have zero probability of being neighbours). Alternatively, if  $L_i$  and  $L_j$  are adjacent, the boundary is likely to have certain characteristics (e.g., a fractal dimension), and the value of the constraint can depend on a measurement.

Equations 4 to 6 show that minimising the Gibbs  $U(\mathbf{w})$  energy for a configuration is equivalent to maximising its probability density function. A preferred embodiment of the invention seeks to find an optimum region label configuration given measurements obtained from the frame  $\mathbf{M}$ , prior knowledge about the labels  $\mathbf{K}$  and the prior probabilities of the labels in the application domain  $\mathbf{Pr}$ . The prior probabilities of the labels are biased by an analysis of contextual data. The problem of optimising the labels over the entire RAG (of the frame) can be solved by iteratively optimising the label at any site,  $i$ . The dependence of the label at region  $i$  on  $\mathbf{M}$ ,  $\mathbf{K}$  and  $\mathbf{Pr}$  is incorporated into the designed clique functions  $V_c(\omega)$ . Therefore the conditional probability density function for  $X_i$  being  $x_i$  at site  $i$  can be written as:

$$P(X_i = x_i | \mathbf{X}, \mathbf{M}, \mathbf{K}, \mathbf{Pr}) = Z_i^{-1} \exp \left[ -\frac{1}{T} \sum_{c \in \mathbf{C}_i} V_c(\omega) \right],$$

$$Z_i = \sum_{x \in L} \exp \left[ -\frac{1}{T} \sum_{c \in C_i} V_c(\omega^x) \right], \quad (7)$$

where  $C_i$  is the subset of  $C$  that consists of cliques that contain  $X_i$  and  $\omega^x$  denotes the configuration which is  $x$  at site  $i$  and agrees with  $\omega$  everywhere else. The prior probabilities of the labels can also be used to bias the initial labels of the sites. For example, labels of previous analysis events can be used to initialise a graph for a later analysis event.

As mentioned above, clique functions can be based on feature measurements from the frame  $M$ , prior knowledge about the labels  $K$ , and prior probabilities of the labels  $Pr$ . Consider, for example, the label "sky" in an Outdoor Scenes application domain. The set of cliques involving region (site)  $i$  on the RAG (i.e.,  $C_i$ ) would typically consist of a unary clique consisting of just region  $i$  and a set of cliques that involve groups of regions, each including region  $i$ , in which each region is a neighbour of each other region in the group.

The unary clique function can be calculated by measuring a collection of features for the region  $i$  and then using these feature measurements as input to a neural network that has been previously trained using examples of sky regions from manually segmented images. Examples of possible features which can be measured for a region include mean R, G and/or B values, mean luminance, variance of the luminance in the region, texture features which may involve measurements derived in the frequency domain, and region shape features such as compactness. The neural network would typically be trained to generate a low value (eg., zero) for regions that have feature measurements that resemble those of the manually segmented sky regions and a high value (eg., 1.0) for those regions that have feature measurements which are very dissimilar to those of the manually segmented regions.

Feature measurements can also be used in clique functions which involve more than one region. For example, the tortuosity of a common boundary between two regions can be used in a clique function involving a pair of regions. The common boundary between a "sky" and a "water" region would typically not be very tortuous whereas the common boundary between "foliage" and "sky" could well be very tortuous.

Prior knowledge can be incorporated into the clique functions in the form of constraints. For example, a clique function involving a "sky" label and a "grass" label might return a high energy value (eg., 1.0) if the region to which the "grass" label is being applied is above the region to which the "sky" label is being applied. In other words, we are using our prior knowledge of the fact that the "sky" regions are usually located above the "grass" regions in frames (except in landscape reflection scenes).

The prior probability of region  $i$  being "sky",  $Pr_{Sky}$ , could also be incorporated into clique functions. One method of doing this is to multiply an existing unary clique function by a factor such as:

$$\left( 1 - \alpha \left( \frac{Pr_{Sky}}{\arg \max_{L \in \mathbf{L}} \mathbf{Pr}_L} \right) \right), \quad (8)$$

where  $\alpha$  is some parameter in the range of (0,1) that weights the contribution of the prior probability to the overall clique function. Prior probabilities can also be incorporated into clique functions involving more than one region. In this case, the multiplying factor for the clique function typically involves the prior probabilities of each of the labels in the clique function.

Equation 7 demonstrates that selecting the most probable label at a site is equivalent to minimising the weighted, by prior probability of the label, Gibbs energy function  $U(\omega)$  at the site. The optimum region label configuration for the frame can be obtained by iteratively visiting each of the  $N$  sites on the graph  $G$  and updating the label at each site. There exist several methods by which the region labels are updated. A new label can be selected for a region from either a uniform distribution of the labels or from the conditional probability distribution of the MRF (ie., the Gibbs Sampler, see Geman and Geman, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6, pp. 721-741, 1984). If more rapid convergence is desirable, then the iterated conditional modes (described by Besag, J., *J. R. Statistical Soc. B*, 48, pp.259-302, 1986) method can be used. In the latter method, sites on the RAG are iteratively visited and, at each site, the label of the region is updated to be the label that has the largest conditional probability distribution. The iterative procedure of visiting and updating sites can be implemented

within a simulated annealing scheme (where the temperature is gradually decreased). The method of updating is not critical for this embodiment of the invention. Instead, it is the inclusion of the prior probability in the calculation of the Gibbs energy  $U(\omega)$  that is a critical.

5

## 6. Contextual Analyser

The contextual analyser 410 in Fig 4 takes the current frame and a ROI 400 for each of the information sources (eg., video signal 110A, and audio signal 110B) and provides information to the adjusting unit 430 on how the prior probabilities of the labels in the application domain 440 should be biased. The function of the contextual analyser 410 as depicted in Fig. 6 has already been discussed in association with the frame event analyser 202 of Fig. 2. A method of adjusting the prior probabilities of labels in an application domain 440 based on the presence of various key words/phrase in the audio signal ROI will be described in more detail hereinafter. Similar methods can be used for other contextual information

15

Each label can be associated with one or more evidence units, where an evidence unit comprises a key word or phrase and a weight factor between 0 and 1. For example, an evidence unit for the label "water" can consist of the key word "beach" and a weighting factor of 0.8. The value of the weighting factor implies the likelihood that the existence of the key word in the audio ROI indicates that "water" is the appropriate label for at least one region in the RAG.

20

Before evidence is collected the sum of the prior probabilities of all labels should sum to 1.0. In other words:

25

$$\sum_{l=1}^L Pr_l = 1.0 \quad (9)$$

As evidence is collected from the ROIs of the contextual information, evidence units are instantiated. The weight factors for the different instantiated evidence units for a given label  $l$ , can be summed to generate the total evidence for the label,  $E_l$ .

30

The  $Pr_l$  values for the labels in the application domain 440 can then be calculated using,

$$Pr_l = (1 + E_l)x, \quad (10)$$

where the value of  $x$  is obtained by solving:

$$\sum_{l=1}^L (1 + E_l)x = 1.0 \quad (11)$$

5           The resulting  $Pr_l$  values can be used directly by the clique functions (see for example Equation 8).

#### 7.       Alternative Embodiments of the Invention

Fig. 12 shows the operation of a video segment analyser 140 according to an  
10   alternative embodiment of the invention in which the video segment analyser 140 is  
integrated with an object-based digital video coding system. In block 250, a first frame in  
a video segment 130 generated by the video segmenter 120 of Fig. 1 is loaded into the  
video segment analyser 140. The frame event analyser 252 receives the loaded frame and  
analyses the frame using the contextual information from the relevant ROI as described  
15   for the frame event analyser 202 in Fig. 2A resulting in a labelled RAG. The labelled  
RAG is then output by the frame event analyser 252 to a region encoder 254 which  
encodes the RAG. The region encoder 254 encodes the regions of the RAG, including  
their adjacency and depth information and semantic labels into a bitstream. In block 256,  
a check is made to determine if the end of the video segment has been reached. If the  
20   checking block 256 returns true (yes), then video segment processing terminates in block  
258. Otherwise, if checking or decision block 256 returns false (no), the next frame in the  
video segment is loaded in block 260.

The motion detector 262 detects motion in the video segment on a frame-by-  
frame basis. It examines any motion, detected from the previous frame, on a region basis.  
25   If the motion of individual regions can be described by a motion model (eg., an affine  
transformation of the region), the model parameters are encoded in the bit stream in block  
266. If the detected motion cannot be described by the motion model then the frame is  
analysed by the frame event analyser 252 and a new RAG is generated and encoded by the  
region encoder 254.

30           In the video segment analyser 140 depicted in Fig. 12, the semantic labels are  
preferably integrated with the coded digital video signal. If the video segment analyser is

integrated with the digital video coding system, the regions can be separately coded in a resolution independent manner. This enables simple reconstruction of a digital video signal at any desired resolution. The method of encoding the digital video signal may be carried out using any of a number of such techniques well known to those skilled in the art. Clearly, the video segment analyser 140 does not necessarily need to be integrated with a digital video coding system. Instead, as noted hereinbefore, the video segment analyser 140 may just generate metadata. In such an embodiment, it is not necessary to process all the video frames in a segment. In other words, only selected frames in a segment need be analysed. It is not the objective of the embodiments of this invention to specify how frames are selected as such selection depends to a large extent on the implementation. For example, a video interpretation system needs to work close to real-time.

Another alternative embodiment of the invention is one in which the video frame segmentation and region labelling process are combined in a single minimisation process.

The foregoing describes only some embodiments of the present invention and modifications, apparent to those skilled in the art, can be made thereto without departing from the scope of the present invention.

In the context of this specification, the word "comprising" means "including principally but not necessarily solely" or "having" or "including" and not "consisting only of". Variations of the word comprising, such as "comprise" and "comprises" have corresponding meanings.